Identifying communities and ranking the drivers' performance in Formula One

Matej Horvat
mh3418@student.uni-lj.si
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113
SI-1000 Ljubljana, Slovenia

Lan Sovinc
ls7312@student.uni-lj.si
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113
SI-1000 Ljubljana, Slovenia

Domen Grzin
dg6781@student.uni-lj.si
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113
SI-1000 Ljubljana, Slovenia

ABSTRACT

This paper explores an approach for ranking the performance of Formula One drivers across different eras using network analysis methods, namely PageRank combined with community detection algorithms. Typically, attempts to rank driver performance have been limited, focusing solely on the evolution of technology and rules in the sport, and attempting to find the best individual driver, ignoring the fact that numerous good drivers never competed directly. To address these challenges, we collected data from all Formula One races since its inception in 1950 through 2022 and used network analysis to create communities of drivers based on their direct competition with each other. Within these communities, we then applied the PageRank algorithm to rank the drivers. Our approach has been shown to produce more meaningful and relevant driver rankings compared to the full graph PageRank results, and effectively recognises the dominance of drivers in their respective eras. Our methodology provides the basis for more sophisticated performance comparisons in sports with long histories and changing conditions.

KEYWORDS

Network Analysis, Formula One, Community Detection, Performance Comparison, PageRank, Leiden Algorithm

1 INTRODUCTION

The FIA Formula One World Championship was introduced in 1950. The word "Formula" in the name refers to the set of rules laid down by the FIA (Fédération Internationale de l'Automobile), to which all cars must conform. Over the years, the championship has been subject to numerous changes in order to keep it at the cutting edge of technology, maintain popularity and spectator numbers, and remain relevant in the world of motorsport.

Technology, engineering and rule changes play a very important role in the sport, but it is often said that the driver makes the biggest difference in the final performance of a constructor. Fans of the sport often compare their favourite drivers to others and try to make them seem better than others, or even hail them as the best ever. Since the beginning of the championship, there have been more than 800 drivers. We believe that the talent and performance of different drivers over the years cannot be objectively compared. Their results can only be compared with the results of those who have raced alongside them. There are some eras in the sport that are not formally defined, which refer to certain technological changes or changes in the rules of the sport. The results of the most successful drivers can often be linked to these eras, but the results still depend too much on the work of engineers and race car designers, rather than on the performance of the individual driver. To solve this problem, we used the principles of network analysis to find an objective way to classify the performance of drivers using the help of Leiden algorithm and node importance classification algorithm.

2 RELATED WORK

Network analytic methods have become a valuable tool in sports research [12]. Their use spans several areas of sports research; in sports management, they can be used to study the relationships between teams and their sponsors [5]. They can help understand the success and performance of teams in competitions [9] or social structures of sports organisations, e.g., what characteristics cause a tennis match to be abandoned [2].

One of the most important methods in network analytics are centrality measures. They provide quantitative metrics for assessing the importance and centrality of nodes in a network. In a network of passes between members of a basketball team, centrality measures rank players according to their importance to the team [6]. We focused our research on PageRank [3]. PageRank was originally developed by Google to rank web pages. It focuses on the idea that important nodes in a network are mainly connected to other important nodes. This is achieved by counting the number and quality of edges of a node, which estimates the importance of the node. Another important method of network analysis is community detection. With these methods, the network is divided into multiple node communities based on criteria such as similarity measures, modularity, or optimization algorithms. In a network of basketball players, players can be grouped based on their performance [4]. Discovering the optimal communities in a network (modularity optimization) is a well-known NP-hard problem [1], so when we talk about community discovery algorithms, we talk about heuristic algorithms. As such, they differ in their speed and correctness of detected communities [8]. We decided to use the **Leiden** modularity optimization algorithm since it gives excellent results and computes them quickly [11].

In recent years, the use of modularity to improve the results of centrality measures has received some attention [7]. Since most real-world networks are modular, this logically leads to better results. There is much work demonstrating the usefulness of such centrality measures, including in social networks, e.g., for identifying influential spreaders during an epidemic.

DOI: https://doi.org/10.26493/scores23.08

3 RESULTS

3.1 Graph without communities

Table 1: Top drivers by Pagerank

Driver	Active years	Titles won	Wins	Podiums
Juan Fangio	1950 – 1958	5	24	35
Graham Hill	1958 – 1975	2	14	36
Jack Brabham	1955 – 1970	3	14	31
Stirling Moss	1951 – 1961	0	16	24
Bruce McLaren	1958 – 1970	0	4	27
Lewis Hamilton	2007–	7	103	192
Alberto Ascari	1950 – 1955	2	13	17
Nino Farina	1950 – 1956	1	5	20
Alain Prost	1980 – 1993	4	51	106
Niki Lauda	1971 – 1985	3	25	54

Looking at the results of running the PageRank algorithm on the graph without communities, we see some very high-performing drivers in the table 1. 8 of the top 10 nodes represent drivers who have won at least one drivers' championship title in their time on the circuit. We can quickly see that many very good drivers are missing from the list (the most prominent example being the legendary Michael Schumacher). The second obvious problem is that we are comparing drivers who raced in completely different eras of the sport. The list includes some of the earliest winners, like Juan Fangio, and some of the most recent, like Lewis Hamilton. The amount of information we can extract from this list is relatively small and is the reason we prefer to analyse PageRank results for each of the communities we detected.

3.2 Detected communities

The Leiden community detection algorithm has identified 6 communities of drivers that have raced with each other the most. Applying the PageRank algorithm to each community yields promising results compared to the data without communities. Results of the PageRank algorithm and detected communities are shown on figure 1.

3.2.1 Community 1. The first community identified contains the drivers who raced from the beginning of Formula One in the 1950s to the 1970s. Of the first 10 top drivers by PageRank, 6 out of 10 drivers became drivers' world champions at least once.

The top 3 drivers from this community were:

- Juan Manuel Fangio: The driver dominated in the early days of Formula One, winning the drivers' championship five times in the 1950s. His record remained undefeated until 2003, when Michael Schumacher became world champion for the sixth time.
- Jack Brabham: Three-time world champion in the late 1950s and 1960s who was also knighted for his racing successes.
- Graham Hill: Two-time world champion in the 1960s and the only driver to win the Triple Crown of Motorsport (the 24 Hours of Le Mans, Indianapolis 500 and Monaco Grand Prix).

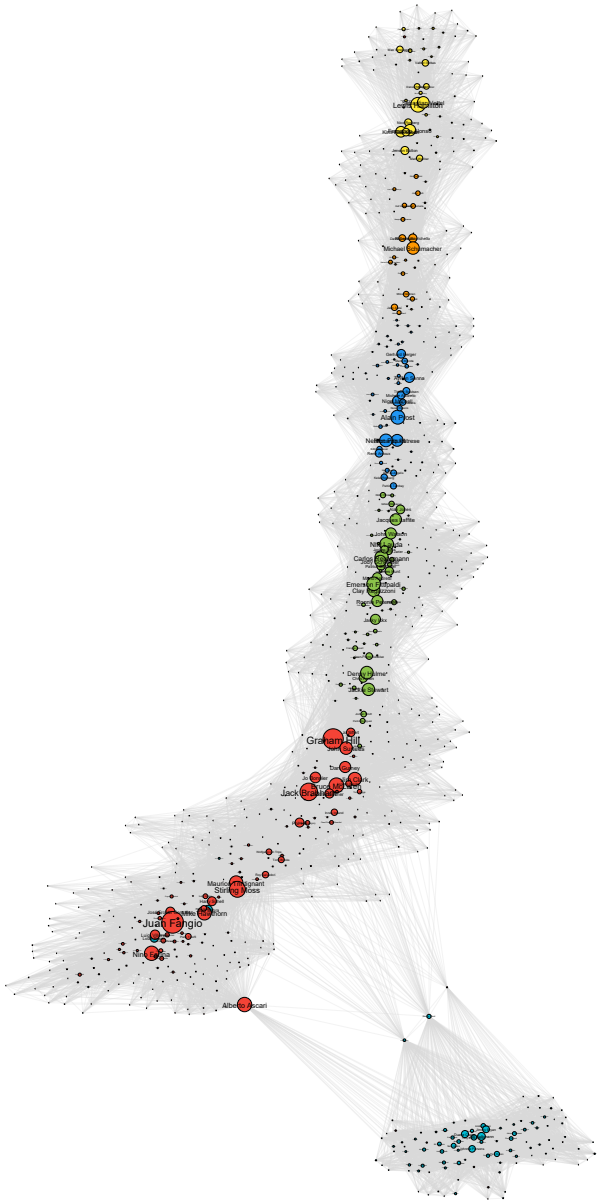

Figure 1: Formula One drivers performance network. The bigger the node, the better the driver's performance. Colored by the detected communities.

The other top finishers include some of the most successful drivers of the era. While not all drivers became world champions, they won many times, scored many points in their careers, and contributed greatly to the sport.

3.2.2 Community 2. The second community contained nodes of drivers that competed from the 1970s to the early 1980s (table 2). 4 of the top 10 nodes represent the drivers, who became world champions. Interestingly, only one of the first three drivers on the list was a world champion, showing that our approach does not

identify the drivers performance only by the number of titles they won. However, it should be noted that the results of the first 5 drivers in the ranking are very similar.

Table 2: Top drivers by Pagerank in community 2

Driver	Active years	Titles won	Wins	Podiums
Emerson Fittipaldi	1970 – 1980	2	14	35
Carlos Reutemann	1972 – 1982	0	12	45
Clay Regazzoni	1970 – 1980	0	5	28
Niki Lauda	1971 – 1985	3	25	54
Jody Scheckter	1972 – 1980	1	10	33
Ronnie Peterson	1970 – 1978	0	10	26
John Watson	1973 – 1985	0	5	20
Jacques Laffite	1974 – 1986	0	6	32
James Hunt	1973 – 1979	1	10	23
Patrick Depailler	1972 – 1980	0	2	19

The top 3 drivers in the list:

- Emerson Fittipaldi: One of the most famous drivers of this era and 2-time champion.
- Carlos Reutemann: Never became world champion, but finished at the top of the championship in the late 1970s and early 1980s, winning 12 Grand Prix.
- Clay Regazzoni: The driver won 5 times, but never became champion. His other achievements include 28 podiums and 209 points.

Also notable are three-time world champion Niki Lauda and one-time world champion James Hunt, who were known for their rivalry in the 1970s.

3.2.3 Community 3. This group consists of drivers from the 1980s to early 1990s and includes many icons of motorsport. In this period there were 6 different world champions and 4 of them are the first 4 on the ranking. The other two were Keke Rosberg and Niki Lauda, which the algorithm assigned to the previous community.

The top finishers were:

- Alain Prost: He was a four-time world champion and also held the record for most Grand Prix wins until 2001.
- Ayrton Senna: One of the sport's greatest legends and a three-time world champion. He is also known for his rivalry with Prost.
- Nelson Piquet: Also won the drivers' championship three times and is considered one of the best drivers of the era.

3.2.4 Community 4. The fourth community is very interesting because it contains drivers from the 1950s and 1960 who competed in Formula One only for the Indianapolis 500 race, which is now exclusively part of the IndyCar Series. The race was on the Formula One calendar from 1950 to 1960. Visualizing the network those drivers stand out by creating a lot of edges between each other and very little between "regular" F1 drivers as they only raced with them once a year. Also, none of those racers ever won the drivers' championship.

The top performers of the community are:

- Jim Rathmann: Winner of the Indianapolis 500 in 1960.
- Duane Carter: Scored a third place in the Indianapolis 500 in 19533.
- Jimmy Bryan: Winner of the Indianapolis 500 in 1958.

3.2.5 Community 5. This community includes drivers from the recent Formula One era (table 3). We have 6 world champions in the top 10 places.

Table 3: Top drivers by Pagerank in community 5

Driver	Active years	Titles won	Wins	Podiums
Lewis Hamilton	2007–	7	103	192
Sebastian Vettel	2007 – 2022	4	53	122
Fernando Alonso	2001–	2	32	102
Kimi Räikkönen	2001 – 2021	1	21	103
Valtteri Bottas	2013–	0	10	67
Max Verstappen	2014–	2	38	82
Nico Rosberg	2006 – 2016	1	23	57
Felipe Massa	2002 – 2017	0	11	41
Daniel Ricciardo	2011 – 2022	0	8	32
Sergio Pérez	2011–	0	6	30

In the top 3 we find:

- Lewis Hamilton: Considered one of the greatest drivers in Formula One, he is a seven-time world champion. He holds many records, including sharing the record for the most drivers' world titles with Schumacher and the most total Grand Prix victories.
- Sebastian Vettel: Four-time consecutive world champion, dominating in the early 2010s.
- Fernando Alonso: The driver with the longest racing experience in the sport and a two-time world champion who regularly scores points in races.

Other important names are Max Verstappen with 2 world titles in 6th place and Valtteri Bottas and Sergio Perez in 5th and 10th place. These two are interesting because they have never been world champions, but have competed for teams that have won many constructors' titles in recent years.

3.2.6 Community 6. The last community identified contains drivers whose careers span from the 1990s to the early 2000s (table 4). There are only 3 world champions in the top 10, but this is mainly due to the dominance of Michael Schumacher.

Table 4: Top drivers by Pagerank in community 6

Driver	Active years	Titles won	Wins	Podiums
Michael Schumacher	1991 – 2012	7	91	55
David Coulthard	1994 – 2008	0	13	62
Rubens Barrichello	1993 – 2011	0	11	68
Mika Häkkinen	1991 – 2001	2	20	51
Ralf Schumacher	1997 – 2007	0	6	27
Giancarlo Fisichella	1996 – 2009	0	3	19
Jacques Villeneuve	1996 – 2006	1	11	23
Jean Alesi	1989 – 2001	0	1	32
Heinz-Harald Frentzen	1994 – 2003	0	3	18
Eddie Irvine	1993 – 2002	0	4	26

The 3 most powerful drivers were:

- Michael Schumacher: For many, Formula One's greatest legend, sharing the record for most world titles won with Lewis Hamilton. He won his first two titles in the nineties, but his dominance on the track was the greatest in the early 2000s, when he won five more in a row.
- David Coulthard: Never won the title, but finished in the top three five times between 1995 and 2001.
- Rubens Barrichello: Also never won the drivers' title, but finished in the top 5 every year between 2000 and 2004.

Other important drivers on the list include two-time world champion Mika Häkkinen and one-time title winner Jacques Villeneuve.

4 METHODS

The research was started by collecting the drivers data. We acquired the open source dataset from the platform Kaggle [10]. It contained data about all the Formula One races, drivers, constructors, qualifying, circuits, lap times, pit stops, championships from 1950 till the present day.

The dataset was composed of 14 tables in the form of .csv text files, where we decided to only include completed championships so we cut out data from the year 2023 onwards. We processed the raw data using Python and the data manipulation and analysis library pandas, using the results of all provided races. For each race we calculated the number of drivers, as some earlier races had multiple drivers racing with the same vehicle.

Then the graphs and networks library NetworkX was used to construct a directed multigraph. We wanted to create a graph that represented who competed with whom, for how long and how good their relative performance was. The graph was composed of nodes, where each node represented a driver. Edges between the drivers represented their ranking in the championship. For each race of every season we made edges based on the following principle: If driver A had a better final position than driver B, then we formed a directed edge from the node representing driver B to A. Each edge has a weight associated with it, depending on the final position in the race of the driver forming the directed edge. We normalised the weights to equalise the effect of each race and prevent some unexpected outliers. This means that the last driver formed directed edges to every other driver that competed in that race, the second last to every other except the last and so forth until the first driver, who did not form any edges.

After creating the graph we continued by detecting communities. For the proof of concept we used the built-in function of NetworkX to detect communities using the Louvain algorithm (nx.community.louvain_communities). We later switched to the Ledien algorithm in the final implementation using the library cdlib.

Once we gathered the detected communities of the drivers who competed together, we continued by classifying driver node performance by using the Pagerank algorithm, for which we used the NetworkX built-in function (nx.pagerank). We also ran the algorithm on the graph without communities to get a baseline performance for each.

To visualise the communities created by the Leiden algorithm, we used the library igraph to create a plot to represent the outcome. We used the Fruchterman-Reingold force-directed graph drawing for the plot, where the node size represents the performance of the node representing the driver and color represents the comunity the driver belongs to.

5 CONCLUSION

Many times, PageRank has proven to be one of the most successful methods of determining the importance of nodes. However, in the case of Formula One, which has been held for decades, the algorithm was insufficient because the result of the calculation was a list of drivers from different eras, and many of the top candidates have never participated in a race. Such ranking of drivers is largely unfounded. We solved the problem by identifying communities. In layman's terms, the approach to grouping drivers into communities is to create communities based on technical improvements or rule changes. Our approach was more systematic, as we calculated communities algorithmically based on the number of races the drivers raced against each other. This allowed us to compare the relative performance of each driver to those they actually raced against. The Leiden algorithm also detected the one special group of racers who only raced at a single event each year, which was not explicitly noted anywhere in the dataset. When we applied PageRank to each of the calculated communities, the results were much more meaningful than the results of the first algorithm run. Each of the communities contains mainly world champion drivers and their close competitors. Most importantly, the drivers in the list all raced together and their careers overlapped to a large extent. This allowed us to create a fair driver ranking, and the amount of knowledge we could extract from the data increased significantly.

In the future, the ranking system could be improved even more. At the moment, we do not compare all drivers directly with each other, but only the drivers in detected communities. After evaluating the performance of the drivers for every community, the analysis could continue. Ideally, a method that compares the relative performance of the top ranked nodes in each community should be developed to compare the drivers in absolute terms.

REFERENCES

[1] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. 2006. On Modularity - NP-Completeness and Beyond.

[2] Kristijan Breznik and Vladimir Batagelj. 2012. Retired Matches Among Male Professional Tennis Players. *Journal of sports science & medicine* 11 (06 2012), 270–8.

[3] S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN* 30, 1-7 (1998), 107–117.

[4] Alessandro Chessa, Pierpaolo D'Urso, Livia De Giovanni, Vincenzina Vitale, and Alfonso Gebbia. 2022. Complex networks for community detection of basketball players. *Annals of Operations Research* (08 2022).

[5] Joe Cobbs. 2011. The dynamics of relationship marketing in international sponsorship networks. *Journal of Business & Industrial Marketing* 26 (10 2011), 590–601.

[6] Jennifer Fewell, Dieter Armbruster, John Ingraham, Alexander Petersen, and James Waters. 2012. Basketball Teams as Strategic Networks. *PloS one* 7 (11 2012), e47445.

[7] Zakariya Ghalmane, Mohammed El Hassouni, Chantal Cherifi, and Hocine Cherifi. 2019. Centrality in modular networks. *EPJ Data Science* 8, 1 (may 2019).

[8] Gnce Keziban Orman, Vincent Labatut, and Hocine Cherifi. 2011. On Accuracy of Community Structure Discovery Algorithms. *Journal of Convergence Information Technology* 6, 11 (nov 2011), 283–292.

[9] Luca Pappalardo and Paolo Cintia. 2017. Quantifying the relation between performance and success in soccer. *Advances in Complex Systems* 21 (05 2017).

[10] Rao R. 2023. Formula 1 World Championship (1950 - 2023). Retrieved from https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020. Accessed on May 9, 2023.

[11] V. A. Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* 9 (2019), 5233.

[12] Hagen Wäsche, Geoff Dickson, Alexander Woll, and Ulrik Brandes. 2017. Social network analysis in sport research: an emerging paradigm. *European Journal for Sport and Society* 14 (05 2017), 1–28.